

Using mixtures in seemingly unrelated linear regression models with non-normal errors

Giuliano Galimberti*, Elena Scardovi, Gabriele Soffritti

Department of Statistical Sciences, University of Bologna

Abstract

Seemingly unrelated linear regression models are introduced in which the distribution of the errors is a finite mixture of Gaussian components. Identifiability conditions are provided. The score vector and the Hessian matrix are derived. Parameter estimation is performed using the maximum likelihood method and an Expectation-Maximisation algorithm is developed. The usefulness of the proposed methods and a numerical evaluation of their properties are illustrated through the analysis of a real dataset.

Keywords: EM algorithm, Gaussian mixture model, Hessian matrix, Score vector.

1. Introduction

“Seemingly unrelated regression equations” is an expression first used by Zellner (1962). It indicates a set of equations for modelling the dependence of D variables ($D \geq 1$) on one or more regressors in which the error terms in the different equations are allowed to be correlated and, thus, the equations should be jointly considered. The range of situations for which models composed of seemingly unrelated regression equations are appropriate is wide, including cross-section data, time-series data and repeated measures (see, e.g., Srivastava and Giles, 1987; Park, 1993).

Seemingly unrelated regression models have been studied through many approaches. In Zellner (1962, 1963) feasible generalized least squares estimators are introduced and their properties are analysed. The maximum likelihood estimator from a Gaussian distribution for the error terms is investigated, for example, in Kmenta and Gilbert (1968); Oberhofer and Kmenta (1974); Magnus (1978); Park (1993). Further developments have been obtained by using bootstrap methods (see, e.g., Rocke, 1989; Rilstone and Veall, 1996) and a likelihood distributional analysis (Fraser *et al.*, 2005). Many studies have been

*Correspondence to: Department of Statistical Sciences, University of Bologna
via Belle Arti 41, 40126 Bologna, Italy. Tel.: +39 051 2098227, Fax: +39 051 232153
Email address: giuliano.galimberti@unibo.it (Giuliano Galimberti)

performed also in a Bayesian framework (see, e.g., Zellner, 1971; Percy, 1992; Ando and Zellner, 2010; Zellner and Ando, 2010a). Most of these methods have been developed under the assumption that the distribution of the error terms is Gaussian. Properties of the feasible generalized least squares estimators under non-Gaussian errors or solutions obtained using other distributions are described, for example, in Srivastava and Maekawa (1995); Kurata (1999); Kowalski *et al.* (1999); Ng (2002); Zellner and Ando (2010b).

The aim of this paper is to propose the use of finite mixtures for modelling the error term distribution in a seemingly unrelated linear regression model. Finite mixture models are widely employed in many areas of multivariate analysis, especially for model-based cluster analysis, discriminant analysis and multivariate density estimation (see, e.g., McLachlan and Peel, 2000). Recently, finite mixtures of Gaussian and Student- t distributions have been employed also in multiple and multivariate linear regression analysis (see, e.g., Bartolucci and Scaccia, 2005; Soffritti and Galimberti, 2011; Galimberti and Soffritti, 2014) to handle non-normal error terms. This approach has the advantage of capturing the effect of omitted nominal regressors from the model and obtaining robust estimates of the regression coefficients when the distribution of the error terms is non-normal. In this paper the same approach is applied to the seemingly unrelated regression model. In particular, the focus is on seemingly unrelated linear regression models in which the error terms are assumed to follow a finite mixture of multivariate Gaussian distributions.

The paper is organized as follows. Section 2 illustrates the theory behind the new methodology. Namely, the novel models are presented in Section 2.1. Theorem 1 provides conditions for the model identifiability (Section 2.2). The score vector and the Hessian matrix for the model parameter are reported in Section 2.3 (Theorems 2 and 3). Details about the maximum likelihood (ML) estimation through an Expectation-Maximisation (EM) algorithm are given in Section 2.4. Results obtained from the analysis of a real dataset using the proposed approach and other methods are presented in Section 3. In Section 4 some concluding remarks are provided. Proofs of Theorems 2 and 3 and other technical results are in Appendix.

2. Seemingly unrelated regression models with a mixture of Gaussian components for the error terms

2.1. The general model

The novel model can be introduced as follows. Let $\mathbf{Y}_i = (Y_{i1}, \dots, Y_{id}, \dots, Y_{iD})'$ be the vector of the D dependent variables for the i th observation, $i = 1, \dots, I$. Furthermore, let \mathbf{x}_{id} be the vector composed of the fixed values of the P_d regressors for the i th observation in the equation for the d th dependent variable, $d = 1, \dots, D$. A seemingly unrelated regression model can be defined through

the following system of equations:

$$\begin{cases} Y_{i1} = \beta_{01} + \mathbf{x}'_{i1}\boldsymbol{\beta}_1 + \epsilon_{i1} \\ \vdots \\ Y_{id} = \beta_{0d} + \mathbf{x}'_{id}\boldsymbol{\beta}_d + \epsilon_{id} \\ \vdots \\ Y_{iD} = \beta_{0D} + \mathbf{x}'_{iD}\boldsymbol{\beta}_D + \epsilon_{iD} \end{cases} \quad i = 1, \dots, I, \quad (1)$$

where β_{0d} , $\boldsymbol{\beta}_d$, and ϵ_{id} are the intercept, the regression coefficient vector and the error term for the i th observation in the equation for the d th dependent variable, respectively. Equation (1) can be written in compact form using the following matrix notation:

$$\mathbf{Y}_i = \boldsymbol{\beta}_0 + \mathbf{X}'_i \boldsymbol{\beta} + \boldsymbol{\epsilon}_i, \quad i = 1, \dots, I, \quad (2)$$

where $\boldsymbol{\beta}_0 = (\beta_{01}, \dots, \beta_{0d}, \dots, \beta_{0D})'$, $\boldsymbol{\beta} = (\boldsymbol{\beta}'_1, \dots, \boldsymbol{\beta}'_d, \dots, \boldsymbol{\beta}'_D)'$, $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, \dots, \epsilon_{id}, \dots, \epsilon_{iD})'$, and \mathbf{X}_i is the following $P \times D$ partitioned matrix:

$$\begin{bmatrix} \mathbf{x}_{i1} & \mathbf{0}_{P_1} & \cdots & \mathbf{0}_{P_1} \\ \mathbf{0}_{P_2} & \mathbf{x}_{i2} & \cdots & \mathbf{0}_{P_2} \\ \vdots & \vdots & & \vdots \\ \mathbf{0}_{P_D} & \mathbf{0}_{P_D} & \cdots & \mathbf{x}_{iD} \end{bmatrix}, \quad (3)$$

with $\mathbf{0}_{P_d}$ denoting the P_d -dimensional null vector and $P = \sum_{d=1}^D P_d$.

Remark 1. Note that this definition of seemingly unrelated regression model differs from the one originally introduced by Zellner (1962); however, these two definitions are equivalent (see, for example, Park (1993)). The choice of the model definition given in equation (2) is motivated by its analytical convenience in deriving some technical results described in this paper.

The proposed model is based on the assumption that the I error terms are independent and identically distributed, and that

$$\boldsymbol{\epsilon}_i \sim \sum_{k=1}^K \pi_k N_D(\boldsymbol{\nu}_k, \boldsymbol{\Sigma}_k), \quad i = 1, \dots, I, \quad (4)$$

where π_k 's are positive weights that sum to 1, the $\boldsymbol{\nu}_k$'s are D -dimensional mean vectors that satisfy the constraint $\sum_{k=1}^K \pi_k \boldsymbol{\nu}_k = \mathbf{0}_D$, the $\boldsymbol{\Sigma}_k$'s are $D \times D$ positive definite symmetric matrices and $N_D(\boldsymbol{\nu}_k, \boldsymbol{\Sigma}_k)$ denotes the D -dimensional Gaussian distribution with parameters $\boldsymbol{\nu}_k$ and $\boldsymbol{\Sigma}_k$.

Given equations (2) and (4), the probability density function (p.d.f.) of the D -dimensional random vector \mathbf{Y}_i is

$$\sum_{k=1}^K \pi_k \phi_D(\mathbf{y}_i; \boldsymbol{\lambda}_k + \mathbf{X}'_i \boldsymbol{\beta}, \boldsymbol{\Sigma}_k), \quad \mathbf{y}_i \in \mathbb{R}^D, \quad i = 1, \dots, I, \quad (5)$$

where $\phi_D(\mathbf{y}_i; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the p.d.f. of the D -dimensional Gaussian distribution $N_D(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ evaluated at \mathbf{y}_i , and $\boldsymbol{\lambda}_k = \boldsymbol{\beta}_0 + \boldsymbol{\nu}_k$. Differently from the $\boldsymbol{\nu}_k$'s, the $\boldsymbol{\lambda}_k$'s are not subject to any constraint. For this reason, in this paper the attention is focused on the vector of the model parameters given by $\boldsymbol{\theta} = (\boldsymbol{\pi}', \boldsymbol{\beta}', \boldsymbol{\theta}'_1, \dots, \boldsymbol{\theta}'_K)'$, where $\boldsymbol{\pi} = (\pi_1, \dots, \pi_{K-1})'$, $\boldsymbol{\theta}_k = (\boldsymbol{\lambda}'_k, \mathbf{v}(\boldsymbol{\Sigma}_k))'$ for $k = 1, \dots, K$, with $\mathbf{v}(\boldsymbol{\Sigma}_k)$ denoting the $\frac{1}{2}D(D+1)$ -dimensional vector formed by stacking the columns of the lower triangular portion of $\boldsymbol{\Sigma}_k$ (see, e.g., Schott, 2005).

Suppose that the i th observation was drawn from the k th component of the mixture. Then, the equation for such an observation would be

$$\mathbf{Y}_i = \boldsymbol{\lambda}_k + \mathbf{X}'_i \boldsymbol{\beta} + \tilde{\boldsymbol{\epsilon}}_{ik}, \quad (6)$$

where $\tilde{\boldsymbol{\epsilon}}_{ik} \sim N_D(\mathbf{0}_D, \boldsymbol{\Sigma}_k)$. The model defined by equation (5) can be seen as a mixture of K seemingly unrelated linear regression models with Gaussian error terms. In this model, observations drawn from different components have different intercepts for the D dependent variables and different covariance matrices for the error terms, while the regression coefficients are equal across components. In the special case where $K = 1$, this model results in the classical seemingly unrelated regression model with Gaussian errors. If $\mathbf{x}_{id} = \mathbf{x}_i \forall d$ (the vectors of the regressors for the D equations coincide), the model proposed by Soffritti and Galimberti (2011) is obtained. Furthermore, the model proposed by Bartolucci and Scaccia (2005) can be obtained when $D = 1$. Finally, if $P_d = 0 \forall d$, model (5) results in the mixture model with K Gaussian components (see, e.g., McLachlan and Peel, 2000).

2.2. Model identifiability

As any finite mixture model, also model (5) is invariant under permutations of the labels of the K components (see, e.g., McLachlan and Peel, 2000). For the proposed model, whose parameter is $\boldsymbol{\theta} = (\boldsymbol{\pi}', \boldsymbol{\beta}', \boldsymbol{\theta}'_1, \dots, \boldsymbol{\theta}'_K)'$, the following theorem holds:

Theorem 1. *The linear regression model (5) is identifiable, provided that, for $d = 1, \dots, D$, vectors $\{\mathbf{x}_{id}, i = 1, \dots, I\}$ do not lie on a common $(P_d - 1)$ -dimensional hyperplane.*

Proof. The identifiability condition described in Theorem 1 is a generalization of the usual condition for the identifiability of a multiple linear regression model. It is required in order to guarantee identifiability of the parameters $\boldsymbol{\beta}$ and $\boldsymbol{\lambda}_1, \dots, \boldsymbol{\lambda}_K$ that characterize the conditional expectations for the D dependent variables.

Furthermore, consider the joint conditional p.d.f. of a random sample $\mathbf{y}_1, \dots, \mathbf{y}_i, \dots, \mathbf{y}_I$ from the model (5), given the fixed values of the regressors contained in $\mathbf{X}_1, \dots, \mathbf{X}_I$:

$$f(\mathbf{y}_1, \dots, \mathbf{y}_I; \mathbf{X}_1, \dots, \mathbf{X}_I, \boldsymbol{\theta}) = \prod_{i=1}^I \left[\sum_{k=1}^K \pi_k \phi_D(\mathbf{y}_i; \boldsymbol{\lambda}_k + \mathbf{X}'_i \boldsymbol{\beta}, \boldsymbol{\Sigma}_k) \right]. \quad (7)$$

It is possible to show that (7) can be written as the following mixture of J Gaussian components:

$$f(\mathbf{y}_1, \dots, \mathbf{y}_I; \mathbf{X}_1, \dots, \mathbf{X}_I, \boldsymbol{\theta}) = \sum_{j=1}^J \pi_j \phi_{D \cdot I}(\mathbf{y}; \boldsymbol{\lambda}_j + \mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Sigma}_j), \quad (8)$$

where $J = K^I$, $\mathbf{y} = (\mathbf{y}'_1, \dots, \mathbf{y}'_i, \dots, \mathbf{y}'_I)'$, $\mathbf{X} = [\mathbf{X}_1 \dots \mathbf{X}_i \dots \mathbf{X}_I]'$, $\pi_j = \prod_{i=1}^I \pi_{k_i^{(j)}}$, $\boldsymbol{\lambda}_j = (\boldsymbol{\lambda}'_{k_1^{(j)}}, \dots, \boldsymbol{\lambda}'_{k_i^{(j)}}, \dots, \boldsymbol{\lambda}'_{k_I^{(j)}})'$, $\boldsymbol{\Sigma}_j = \text{diag}(\boldsymbol{\Sigma}_{k_1^{(j)}}, \dots, \boldsymbol{\Sigma}_{k_i^{(j)}}, \dots, \boldsymbol{\Sigma}_{k_I^{(j)}})$ is a block diagonal matrix, and $\mathbf{k}^{(j)} = (k_1^{(j)}, \dots, k_I^{(j)})'$ is the j th element of the set $A_{K,I} = \{(k_1, \dots, k_I)' : k_i \in \{1, \dots, K\}, i = 1, \dots, I\}$ containing the J arrangements of the first K positive integers amongst I with repetitions. The proof can be completed by showing that mixtures (8) are identifiable. The proof of this latter result can be found in Soffritti and Galimberti (2011). \square

2.3. Score vector and Hessian matrix

Given a random sample $\mathbf{y}_1, \dots, \mathbf{y}_i, \dots, \mathbf{y}_I$ from the model (5), the log-likelihood is

$$l(\boldsymbol{\theta}) = \sum_{i=1}^I \ln \left(\sum_{k=1}^K \pi_k \phi_D(\mathbf{y}_i; \boldsymbol{\lambda}_k + \mathbf{X}'_i \boldsymbol{\beta}, \boldsymbol{\Sigma}_k) \right). \quad (9)$$

The log-likelihood (9) can be used to derive the ML estimator of $\boldsymbol{\theta}$. Furthermore, Redner and Walker (1984) showed that, under suitable conditions, an estimate of the asymptotic variance of the ML estimator of the parameters in a finite mixture model can be obtained using the Hessian matrix. In order to obtain the score vector and the Hessian matrix the following notation is introduced. Let

$$f_{ki} = \frac{\pi_k}{(2\pi)^{D/2} \det(\boldsymbol{\Sigma}_k)^{1/2}} \exp \left[-\frac{1}{2} (\mathbf{y}_i - \boldsymbol{\lambda}_k - \mathbf{X}'_i \boldsymbol{\beta})' \boldsymbol{\Sigma}_k^{-1} (\mathbf{y}_i - \boldsymbol{\lambda}_k - \mathbf{X}'_i \boldsymbol{\beta}) \right];$$

$\alpha_{ki} = \frac{f_{ki}}{(\sum_{i=1}^K f_{ki})}$; $\mathbf{a}_k = \frac{1}{\pi_k} \mathbf{e}_k$ for $k = 1, \dots, K-1$ and $\mathbf{a}_K = -\frac{1}{\pi_K} \mathbf{1}_{(K-1)}$, where \mathbf{e}_k is the k th column of $\mathbf{I}_{(K-1)}$ (the identity matrix of order $K-1$) and $\mathbf{1}_{(K-1)}$ denotes the $(K-1)$ -dimensional vector having each component equal to 1; $\mathbf{b}_{ki} = \boldsymbol{\Sigma}_k^{-1} (\mathbf{y}_i - \boldsymbol{\lambda}_k - \mathbf{X}'_i \boldsymbol{\beta})$; $\mathbf{B}_{ki} = \boldsymbol{\Sigma}_k^{-1} - \mathbf{b}_{ki} \mathbf{b}'_{ki}$;

$$\mathbf{c}_{ki} = \begin{bmatrix} \mathbf{b}_{ki} \\ -\frac{1}{2} \mathbf{G}' \text{vec}(\mathbf{B}_{ki}) \end{bmatrix},$$

where \mathbf{G} denotes the duplication matrix and $\text{vec}(\mathbf{B}_{ki})$ denotes the vector formed by stacking the columns of the matrix \mathbf{B}_{ki} one underneath the other (see, e.g., Schott, 2005).

Theorem 2. The score vector for the parameters of model (5) is composed of the sub-vectors $\frac{\partial}{\partial \boldsymbol{\pi}'} l(\boldsymbol{\theta})$, $\frac{\partial}{\partial \boldsymbol{\beta}'} l(\boldsymbol{\theta})$, $\frac{\partial}{\partial \boldsymbol{\theta}_1'} l(\boldsymbol{\theta})$, \dots , $\frac{\partial}{\partial \boldsymbol{\theta}_K'} l(\boldsymbol{\theta})$, where

$$\begin{aligned}\frac{\partial}{\partial \boldsymbol{\pi}} l(\boldsymbol{\theta}) &= \sum_{i=1}^I \bar{\mathbf{a}}_i, \\ \frac{\partial}{\partial \boldsymbol{\beta}} l(\boldsymbol{\theta}) &= \sum_{i=1}^I \mathbf{X}_i \bar{\mathbf{b}}_i, \\ \frac{\partial}{\partial \boldsymbol{\theta}_k} l(\boldsymbol{\theta}) &= \sum_{i=1}^I \alpha_{ki} \mathbf{c}_{ki}, \quad k = 1, \dots, K,\end{aligned}$$

with $\bar{\mathbf{a}}_i = \sum_{k=1}^K \alpha_{ki} \mathbf{a}_k$ and $\bar{\mathbf{b}}_i = \sum_{k=1}^K \alpha_{ki} \mathbf{b}_{ki}$.

Theorem 3. The Hessian matrix $H(\boldsymbol{\theta})$ for the parameters of model (5) is equal to

$$\begin{bmatrix} \frac{\partial^2}{\partial \boldsymbol{\pi} \partial \boldsymbol{\pi}'} l(\boldsymbol{\theta}) & \frac{\partial^2}{\partial \boldsymbol{\pi} \partial \boldsymbol{\beta}'} l(\boldsymbol{\theta}) & \frac{\partial^2}{\partial \boldsymbol{\pi} \partial \boldsymbol{\theta}_1'} l(\boldsymbol{\theta}) & \dots & \frac{\partial^2}{\partial \boldsymbol{\pi} \partial \boldsymbol{\theta}_K'} l(\boldsymbol{\theta}) \\ \frac{\partial^2}{\partial \boldsymbol{\beta} \partial \boldsymbol{\pi}'} l(\boldsymbol{\theta}) & \frac{\partial^2}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} l(\boldsymbol{\theta}) & \frac{\partial^2}{\partial \boldsymbol{\beta} \partial \boldsymbol{\theta}_1'} l(\boldsymbol{\theta}) & \dots & \frac{\partial^2}{\partial \boldsymbol{\beta} \partial \boldsymbol{\theta}_K'} l(\boldsymbol{\theta}) \\ \frac{\partial^2}{\partial \boldsymbol{\theta}_1 \partial \boldsymbol{\pi}'} l(\boldsymbol{\theta}) & \frac{\partial^2}{\partial \boldsymbol{\theta}_1 \partial \boldsymbol{\beta}'} l(\boldsymbol{\theta}) & \frac{\partial^2}{\partial \boldsymbol{\theta}_1 \partial \boldsymbol{\theta}_1'} l(\boldsymbol{\theta}) & \dots & \frac{\partial^2}{\partial \boldsymbol{\theta}_1 \partial \boldsymbol{\theta}_K'} l(\boldsymbol{\theta}) \\ \dots & \dots & \dots & \dots & \dots \\ \frac{\partial^2}{\partial \boldsymbol{\theta}_K \partial \boldsymbol{\pi}'} l(\boldsymbol{\theta}) & \frac{\partial^2}{\partial \boldsymbol{\theta}_K \partial \boldsymbol{\beta}'} l(\boldsymbol{\theta}) & \frac{\partial^2}{\partial \boldsymbol{\theta}_K \partial \boldsymbol{\theta}_1'} l(\boldsymbol{\theta}) & \dots & \frac{\partial^2}{\partial \boldsymbol{\theta}_K \partial \boldsymbol{\theta}_K'} l(\boldsymbol{\theta}) \end{bmatrix}, \quad (10)$$

where

$$\begin{aligned}\frac{\partial^2}{\partial \boldsymbol{\pi} \partial \boldsymbol{\pi}'} l(\boldsymbol{\theta}) &= - \sum_{i=1}^I \bar{\mathbf{a}}_i \bar{\mathbf{a}}_i', \\ \frac{\partial^2}{\partial \boldsymbol{\pi} \partial \boldsymbol{\beta}'} l(\boldsymbol{\theta}) &= \sum_{i=1}^I \left[\left(\sum_{k=1}^K \alpha_{ki} \mathbf{a}_k \mathbf{b}_{ki}' \right) - \bar{\mathbf{a}}_i \bar{\mathbf{b}}_i' \right] \mathbf{X}_i', \\ \frac{\partial^2}{\partial \boldsymbol{\pi} \partial \boldsymbol{\theta}_k'} l(\boldsymbol{\theta}) &= \sum_{i=1}^I \alpha_{ki} (\mathbf{a}_k - \bar{\mathbf{a}}_i) \mathbf{c}_{ki}', \quad k = 1, \dots, K, \\ \frac{\partial^2}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} l(\boldsymbol{\theta}) &= - \sum_{i=1}^I \mathbf{X}_i [\bar{\mathbf{B}}_i + \bar{\mathbf{b}}_i \bar{\mathbf{b}}_i'] \mathbf{X}_i', \\ \frac{\partial^2}{\partial \boldsymbol{\beta} \partial \boldsymbol{\theta}_k'} l(\boldsymbol{\theta}) &= - \sum_{i=1}^I \alpha_{ki} \mathbf{X}_i [\mathbf{F}_{ki} - (\mathbf{b}_{ki} - \bar{\mathbf{b}}_i) \mathbf{c}_{ki}'], \quad k = 1, \dots, K, \\ \frac{\partial^2}{\partial \boldsymbol{\theta}_k \partial \boldsymbol{\theta}_k'} l(\boldsymbol{\theta}) &= - \sum_{i=1}^I \alpha_{ki} [\mathbf{C}_{ki} - (1 - \alpha_{ki}) \mathbf{c}_{ki} \mathbf{c}_{ki}'], \quad k = 1, \dots, K, \\ \frac{\partial^2}{\partial \boldsymbol{\theta}_k \partial \boldsymbol{\theta}_h'} l(\boldsymbol{\theta}) &= - \sum_{i=1}^I \alpha_{ki} \alpha_{hi} \mathbf{c}_{ki} \mathbf{c}_{hi}', \quad \forall k \neq h,\end{aligned}$$

with $\bar{\mathbf{B}}_i = \sum_{k=1}^K \alpha_{ki} (\Sigma_k^{-1} - \mathbf{b}_{ki} \mathbf{b}_{ki}')$, $\mathbf{F}_{ki} = \begin{bmatrix} \Sigma_k^{-1} & (\mathbf{b}_{ki}' \otimes \Sigma_k^{-1}) \mathbf{G} \end{bmatrix}$ and

$$\mathbf{C}_{ki} = \begin{bmatrix} \Sigma_k^{-1} & (\mathbf{b}_{ki}' \otimes \Sigma_k^{-1}) \mathbf{G} \\ \mathbf{G}' (\mathbf{b}_{ki} \otimes \Sigma_k^{-1}) & \frac{1}{2} \mathbf{G}' [(\Sigma_k^{-1} - 2\mathbf{B}_{ki}) \otimes \Sigma_k^{-1}] \mathbf{G} \end{bmatrix}.$$

Proofs of Theorems 2 and 3 are provided in Appendix A and Appendix B, respectively.

Remark 2. Theorems 2 and 3 provide the score vector and the Hessian matrix not only for the model proposed in this paper, but also for the models introduced in Bartolucci and Scaccia (2005) and Soffritti and Galimberti (2011), after some suitable simplifications. Furthermore, they represent a generalization of Theorem 1 in Boldea and Magnus (2009).

2.4. An EM algorithm for maximum likelihood estimation

The score vector and the Hessian matrix described in Section 2.3 can be used to compute the ML estimates of the model parameter $\boldsymbol{\theta}$ through a Newton-Raphson algorithm for the maximisation of $l(\boldsymbol{\theta})$ in equation (9). However, the evaluation of the Hessian matrix at each iteration can be computationally expensive, especially with large samples. In order to avoid this problem, in this Section an EM algorithm is developed, using the approach for incomplete-data problems (Dempster *et al.*, 1977; McLachlan and Krishnan, 2008). This approach is widely employed in finite mixture models, where the source of unobservable information is the specific component of the mixture model that generates each sample observation. Specifically, this unobservable information for the i th observation can be described by the K -dimensional vector $\mathbf{z}_i' = (z_{i1}, \dots, z_{iK})$, where $z_{ik} = 1$ when \mathbf{y}_i is generated from the k th component, and $z_{ik} = 0$ otherwise, for $k = 1, \dots, K$. Thus, $\sum_{k=1}^K z_{ik} = 1$, $i = 1, \dots, I$.

Consider the following hierarchical representation for $\mathbf{y}_i | \mathbf{X}_i$:

$$\mathbf{z}_i \sim \text{mult}(1, \pi_1, \dots, \pi_K),$$

$$\mathbf{y}_i | (\mathbf{X}_i, z_{ik} = 1) \sim N_D(\boldsymbol{\lambda}_k + \mathbf{X}_i' \boldsymbol{\beta}, \Sigma_k),$$

where $\text{mult}(1, \pi_1, \dots, \pi_K)$ denotes the K -dimensional multinomial distribution with parameters π_1, \dots, π_K , and assume that this representation independently holds for $i = 1, \dots, I$. Then, the complete-data log-likelihood $l_c(\boldsymbol{\theta})$ of model (5) can be expressed as

$$l_c(\boldsymbol{\theta}) = \sum_{i=1}^I \sum_{k=1}^K z_{ik} \ln f_{ki}. \quad (11)$$

The first order differential of $l_c(\boldsymbol{\theta})$ is

$$\begin{aligned}
dl_c(\boldsymbol{\theta}) &= \sum_{i=1}^I \sum_{k=1}^K z_{ik} d \ln f_{ki} \\
&= (d\boldsymbol{\pi})' \sum_{k=1}^K z_{\cdot k} \mathbf{a}_k + (d\boldsymbol{\beta})' \sum_{i=1}^I \sum_{k=1}^K z_{ik} \mathbf{X}_i \mathbf{b}_{ki} + \sum_{k=1}^K (d\boldsymbol{\theta}_k)' \sum_{i=1}^I z_{ik} \mathbf{c}_{ki} \\
&= (d\boldsymbol{\pi})' \sum_{k=1}^K z_{\cdot k} \mathbf{a}_k \\
&\quad + \sum_{i=1}^I \sum_{k=1}^K z_{ik} [(d\boldsymbol{\lambda}_k)' + (d\boldsymbol{\beta})' \mathbf{X}_i] \mathbf{b}_{ki}
\end{aligned} \tag{12}$$

$$- \frac{1}{2} \sum_{k=1}^K d(\mathbf{v} \boldsymbol{\Sigma}_k)' \mathbf{G}' \text{vec} \left(\sum_{i=1}^I z_{ik} \mathbf{B}_{ki} \right) \tag{13}$$

where the second and third equalities are obtained using equation (A.11) in Appendix A, and $z_{\cdot k} = \sum_{i=1}^I z_{ik}$.

To determine the solution of each M step of the EM algorithm, it is convenient to introduce the following notation. Let dl_{c2} and dl_{c3} denote the expressions in equations (12) and (13), respectively. Let $\boldsymbol{\gamma} = (\boldsymbol{\lambda}'_1, \dots, \boldsymbol{\lambda}'_K, \boldsymbol{\beta}')'$ be the $(D \cdot K + P)$ -dimensional vector comprising the intercepts of all components and regression coefficients for all dependent variables. \mathbf{O}_k is a matrix of dimension $(D \cdot K) \times D$ obtained extracting the columns of the matrix $\mathbf{I}_{(D \cdot K)}$ from the $(1 + (k-1) \cdot D)$ th to the $(D + (k-1) \cdot D)$ th, for $k = 1, \dots, K$. Furthermore, let $\mathbf{X}_{ki} = \begin{bmatrix} \mathbf{O}_k \\ \mathbf{X}_i \end{bmatrix}$; this is a matrix of dimension $(D \cdot K + P) \times D$ such that $\mathbf{X}'_{ki} \boldsymbol{\gamma} = \boldsymbol{\lambda}_k + \mathbf{X}'_i \boldsymbol{\beta}$ and $\mathbf{X}'_{ki} d\boldsymbol{\gamma} = d\boldsymbol{\lambda}_k + \mathbf{X}'_i d\boldsymbol{\beta}$. Using this latter notation, the expressions of dl_{c2} and dl_{c3} in equations (12) and (13) turn into

$$\begin{aligned}
dl_{c2} &= \sum_{i=1}^I \sum_{k=1}^K z_{ik} (d\boldsymbol{\gamma})' \mathbf{X}_{ki} \mathbf{b}_{ki} \\
&= (d\boldsymbol{\gamma})' \sum_{i=1}^I \sum_{k=1}^K z_{ik} \mathbf{X}_{ki} \boldsymbol{\Sigma}_k^{-1} (\mathbf{y}_i - \mathbf{X}'_{ki} \boldsymbol{\gamma}) \\
&= (d\boldsymbol{\gamma})' \sum_{i=1}^I \sum_{k=1}^K z_{ik} \mathbf{X}_{ki} \boldsymbol{\Sigma}_k^{-1} \mathbf{y}_i \\
&\quad - (d\boldsymbol{\gamma})' \left(\sum_{i=1}^I \sum_{k=1}^K z_{ik} \mathbf{X}_{ki} \boldsymbol{\Sigma}_k^{-1} \mathbf{X}'_{ki} \right) \boldsymbol{\gamma},
\end{aligned} \tag{14}$$

$$\begin{aligned}
dl_{c3} &= -\frac{1}{2} \sum_{k=1}^K d(\mathbf{v}\boldsymbol{\Sigma}_k)' \mathbf{G}^\top \text{vec} \left(\sum_{i=1}^I z_{ik} \boldsymbol{\Sigma}_k^{-1} - \sum_{i=1}^I z_{ik} \mathbf{b}_{ki} \mathbf{b}_{ki}' \right) \\
&= -\frac{1}{2} \sum_{k=1}^K d(\mathbf{v}\boldsymbol{\Sigma}_k)' \mathbf{G}' \text{vec} (z_{\cdot k} \boldsymbol{\Sigma}_k^{-1} - \boldsymbol{\Sigma}_k^{-1} \mathbf{S}_k \boldsymbol{\Sigma}_k^{-1}), \tag{15}
\end{aligned}$$

where $\mathbf{S}_k = \sum_{i=1}^I z_{ik} (\mathbf{y}_i - \mathbf{X}_{ki}' \boldsymbol{\gamma}) (\mathbf{y}_i - \mathbf{X}_{ki}' \boldsymbol{\gamma})'$. Using equation (15) and some properties of the vec operator (see, in particular, Schott, 2005, Theorem 8.11), it is also possible to write

$$\begin{aligned}
dl_{c3} &= \frac{1}{2} \sum_{k=1}^K d(\mathbf{v}\boldsymbol{\Sigma}_k)' \mathbf{G}' \text{vec} [\boldsymbol{\Sigma}_k^{-1} (\mathbf{S}_k - z_{\cdot k} \boldsymbol{\Sigma}_k) \boldsymbol{\Sigma}_k^{-1}] \\
&= \frac{1}{2} \sum_{k=1}^K d(\mathbf{v}\boldsymbol{\Sigma}_k)' \mathbf{G}' (\boldsymbol{\Sigma}_k^{-1} \otimes \boldsymbol{\Sigma}_k^{-1}) \mathbf{G} \mathbf{v} (\mathbf{S}_k - z_{\cdot k} \boldsymbol{\Sigma}_k). \tag{16}
\end{aligned}$$

Thus, the following alternative expression for $dl_c(\boldsymbol{\theta})$ holds:

$$\begin{aligned}
dl_c(\boldsymbol{\theta}) &= (d\boldsymbol{\pi})' \sum_{k=1}^K z_{\cdot k} \mathbf{a}_k + (d\boldsymbol{\gamma})' \sum_{i=1}^I \sum_{k=1}^K z_{ik} \mathbf{X}_{ki} \boldsymbol{\Sigma}_k^{-1} \mathbf{y}_i \\
&\quad - (d\boldsymbol{\gamma})' \left(\sum_{i=1}^I \sum_{k=1}^K z_{ik} \mathbf{X}_{ki} \boldsymbol{\Sigma}_k^{-1} \mathbf{X}_{ki}' \right) \boldsymbol{\gamma} \\
&\quad + \frac{1}{2} \sum_{k=1}^K d(\mathbf{v}\boldsymbol{\Sigma}_k)' \mathbf{G}' (\boldsymbol{\Sigma}_k^{-1} \otimes \boldsymbol{\Sigma}_k^{-1}) \mathbf{G} \mathbf{v} (\mathbf{S}_k - z_{\cdot k} \boldsymbol{\Sigma}_k). \tag{17}
\end{aligned}$$

The first derivatives of $l_c(\boldsymbol{\theta})$ with respect to the parameters $\boldsymbol{\pi}$, $\boldsymbol{\gamma}$ and $\mathbf{v}\boldsymbol{\Sigma}_k$ ($k = 1, \dots, K$) are:

$$\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\pi}} l_c(\boldsymbol{\theta}) &= \sum_{k=1}^K z_{\cdot k} \mathbf{a}_k, \\
\frac{\partial}{\partial \boldsymbol{\gamma}} l_c(\boldsymbol{\theta}) &= \sum_{i=1}^I \sum_{k=1}^K z_{ik} \mathbf{X}_{ki} \boldsymbol{\Sigma}_k^{-1} \mathbf{y}_i - \left(\sum_{i=1}^I \sum_{k=1}^K z_{ik} \mathbf{X}_{ki} \boldsymbol{\Sigma}_k^{-1} \mathbf{X}_{ki}' \right) \boldsymbol{\gamma}, \\
\frac{\partial}{\partial (\mathbf{v}\boldsymbol{\Sigma}_k)} l_c(\boldsymbol{\theta}) &= \frac{1}{2} \mathbf{G}' (\boldsymbol{\Sigma}_k^{-1} \otimes \boldsymbol{\Sigma}_k^{-1}) \mathbf{G} \mathbf{v} (\mathbf{S}_k - z_{\cdot k} \boldsymbol{\Sigma}_k), \quad k = 1, \dots, K.
\end{aligned}$$

In order to maximise $l_c(\boldsymbol{\theta})$ these derivatives are set equal to zero. By solving the resulting system of equations the following expressions are obtained:

$$\pi_k^* = z_{\cdot k} / I, \quad k = 1, \dots, K, \tag{18}$$

and, provided that the matrix $\sum_{i=1}^I \sum_{k=1}^K z_{ik} \mathbf{X}_{ki} \boldsymbol{\Sigma}_k^{-1} \mathbf{X}_{ki}'$ is non-singular,

$$\boldsymbol{\gamma}^* = \left(\sum_{i=1}^I \sum_{k=1}^K z_{ik} \mathbf{X}_{ki} \boldsymbol{\Sigma}_k^{-1} \mathbf{X}_{ki}' \right)^{-1} \sum_{i=1}^I \sum_{k=1}^K z_{ik} \mathbf{X}_{ki} \boldsymbol{\Sigma}_k^{-1} \mathbf{y}_i \tag{19}$$

$$\boldsymbol{\Sigma}_k^* = z_{\cdot k}^{-1} \mathbf{S}_k, \quad k = 1, \dots, K. \tag{20}$$

Using some initial value for θ , say $\theta^{(0)}$, the E-step on the $(r + 1)$ th iteration of the EM algorithm is effected by simply replacing z_{ik} by $E_{\theta^{(r)}}(z_{ik}|\mathbf{y}_i, \mathbf{x}_i) = Pr_{\theta^{(r)}}(z_{ik} = 1|\mathbf{y}_i, \mathbf{x}_i) = p_{ik}^{(r)}$, which is the posterior probability that \mathbf{y}_i is generated from the k th component of the mixture. Namely:

$$p_{ik}^{(r)} = \frac{\pi_k^{(r)} \phi_D(\mathbf{y}_i; \boldsymbol{\lambda}_k^{(r)} + \mathbf{X}_i' \boldsymbol{\beta}^{(r)}, \boldsymbol{\Sigma}_k^{(r)})}{\sum_{h=1}^K \pi_h^{(r)} \phi_D(\mathbf{y}_i; \boldsymbol{\lambda}_h^{(r)} + \mathbf{X}_i' \boldsymbol{\beta}^{(r)}, \boldsymbol{\Sigma}_h^{(r)})}.$$

On the M-step at the $(r + 1)$ th iteration of the EM algorithm, the updated estimates of the model parameters $\pi_k^{(r+1)}$, $\boldsymbol{\gamma}^{(r+1)}$ and $\boldsymbol{\Sigma}_k^{(r+1)}$ are computed using equations (18), (19) and (20), respectively, where z_{ik} is replaced by $p_{ik}^{(r)}$. As equation (19) depends on the $\boldsymbol{\Sigma}_k$'s and equation (20) depends on $\boldsymbol{\gamma}$, the updated estimates of such parameters at the $(r + 1)$ th iteration are obtained through an iterative process in which the estimate of $\boldsymbol{\gamma}$ is updated, given an estimate of the $\boldsymbol{\Sigma}_k$'s, and vice versa, until convergence. As far as the choice of $\theta^{(0)}$ is concerned, several strategies can be used (see, e.g., Galimberti and Soffritti, 2014). For example, multiple random initializations can be considered. Otherwise, $\boldsymbol{\beta}^{(0)}$ can be obtained by fitting the standard seemingly unrelated regression model; the sample residuals of this model can be used to derive starting values for the remaining parameters, for example by using them to fit a Gaussian mixture model.

3. Experimental results

The usefulness and effectiveness of the methods described in Section 2 are illustrated through the analysis of the Australian Institute of Sport (AIS) dataset (Cook and Weisberg, 1994). Namely, the interest is focused on studying the joint linear dependence of four biometrical variables (body mass index (BMI), sum of skin folds (SSF), percentage of body fat (PBF), lean body mass (LBM)) on three variables providing information about blood composition (red cell count (RCC), white cell count (WCC), plasma ferritine concentration (PFC)). The same problem was investigated by Soffritti and Galimberti (2011) using multi-variate linear regression models.

A first study is performed to select the regressors to be used for each biometrical variable in a seemingly unrelated linear regression model given by equation (5). The main results are summarized in Section 3.1. Properties of the ML estimates of the regression coefficients for the selected model are numerically evaluated (see Section 3.2). All analyses are performed in the R environment (R Core Team, 2013). A specific function implementing the ML estimation through the EM algorithm and the calculation of the Hessian matrix is used. The starting values of the model parameters are obtained through a strategy that fits Gaussian mixture models to the sample residuals of the classical seemingly unrelated linear regression model. The EM algorithm is stopped when the number of iterations reaches 500 or $|l_{\infty}^{(r+1)} - l^{(r)}| < 10^{-8}$, where $l^{(r)}$ is the

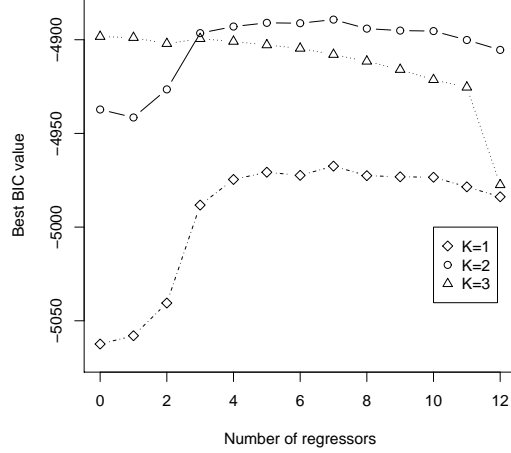


Figure 1: Best BIC values by total number of regressors and number of components.

log-likelihood value from iteration r , and $l_{\infty}^{(r+1)}$ is the asymptotic estimate of the log-likelihood at iteration $r + 1$ (McNicholas and Murphy, 2008). The stopping rules for each M step are either when the mean Euclidean distance between two consecutive estimated vectors of the model parameters is lower than 10^{-8} or when the number of iterations reaches the maximum of 500.

3.1. Selection of the regressors

Seemingly unrelated linear regression models from equation (5) are estimated for $K = 1, 2, 3$. For each of these values, an exhaustive search is performed to select the relevant regressors for each of the $D = 4$ dependent variables. Namely, for each value of K , $2^{3 \cdot D} = 4096$ different regression models are fitted to the dataset, thus resulting in 12288 different seemingly unrelated linear regression models. The total number P of regressors included in a model is between 0 and 12.

The EM algorithm has failed due to the singularity of some matrices for two models when $K = 2$ and 40 models when $K = 3$. The choice of the best model among the estimated ones is performed using the Bayesian Information Criterion (Schwarz, 1978):

$$BIC_M = 2 \max [l_M] - \text{npar}_M \log(I),$$

where $\max [l_M]$ is the maximum of the log-likelihood of a model M for the given sample of I observations, and npar_M is the number of unconstrained parameters to be estimated for that model. This criterion allows to trade-off the fit and parsimony of a given model: the greater the BIC , the better the model.

Table 1: Maximized log-likelihood and BIC value for the best models with K components ($K = 1, 2, 3$) fitted to the AIS dataset.

K	P	$l_M(\hat{\theta})$	npar_M	BIC_M
1	7	-2427.993	21	-4967.46
2	7	-2349.083	36	-4889.26
3	0	-2332.382	44	-4898.33

Figure 1 shows the BIC values of the fitted models with the best trade-off (i.e., the maximum value of the BIC) among all the models having the same values of K and P , for $K = 1, 2, 3$ and $P = 0, \dots, 12$. By comparing models having the same value of P it emerges that the best performance is obtained using models with three components when the total number of regressors is low ($P = 0, 1, 2$); otherwise, models with two components should be preferred. Thus, the introduction of a finite mixture for the distribution of the error terms allows to obtain a relevant improvement with respect to the classical seemingly unrelated regression model with Gaussian errors, for all P .

If models are compared by controlling the number of components, $P = 7$ regressors should be used when $K = 1, 2$. Namely, for both values of K , the selected regressors for the equations of the variables BMI, PBF and LBM are RCC and PFC; only RCC is selected as a relevant regressor for the equation of SSF. When $K = 3$, the best trade-off is obtained using a model without regressors. Some results concerning these three latter models are illustrated in Table 1. Overall, according to the BIC the best model is the one with $K = 2$. In this model, the estimates of the parameters π_1 and π_2 are 0.619 and 0.381. Tables 2 and 3 report the estimates of the remaining parameters. Compared to the second component, the first component is characterized by lower values of the intercepts for all dependent variables and lower variances for BMI, SSF and PBF. Further differences between components concern some correlations (see the lower triangular parts of $\hat{\Sigma}_1$ and $\hat{\Sigma}_2$ in Table 2). The estimated standard errors of the ML estimators of the regression coefficients in Table 3 are computed as the square root of the diagonal elements of $H(\hat{\theta})^{-1}$ that refer to β . The asymptotic confidence intervals for the regression coefficients in Table 3 are obtained by assuming an asymptotic normal distribution for the ML estimators. None of such intervals contains the 0 value.

The best model can be used to assign each athlete to the component of the mixture that registered the highest posterior probability, thus producing a partition of the sample into two clusters. Most of the athletes assigned to the second cluster are female (79.2%), while 68.8% of the athletes classified in the first cluster are male (Tab. 5). This classification of the athletes is statistically associated with athletes' gender ($\chi^2 = 43.96$, $p\text{-value} = 3.36 \cdot 10^{-11}$). Thus, the omitted regressor captured by the selected model has an effect which is strongly connected with athletes' gender.

Table 2: Estimates of parameters λ_k and Σ_k obtained from the best model fitted to the AIS dataset. Estimated correlation coefficients between dependent variables (in italics) are reported in the lower triangular parts of $\hat{\Sigma}_1$ and $\hat{\Sigma}_2$.

	BMI	SSF	PBF	LBM
$\hat{\lambda}'_1$	10.04	86.57	23.19	-7.02
$\hat{\lambda}'_2$	12.99	136.43	32.52	-4.88
$\hat{\Sigma}_1$	3.96 <i>0.198</i> <i>-0.017</i> <i>0.810</i>	5.14 169.94 <i>0.899</i> <i>0.017</i>	-0.09 31.21 7.10 <i>-0.278</i>	18.99 2.63 -8.73 138.82
$\hat{\Sigma}_2$	6.85 <i>0.244</i> <i>0.080</i> <i>0.681</i>	17.43 744.38 <i>0.928</i> <i>-0.244</i>	0.89 107.03 17.88 <i>-0.435</i>	14.59 -54.50 -15.05 67.07

Table 3: Estimates of the regression coefficients (r.c.) calculated from the best model fitted to the AIS dataset and their estimated standard errors (s.e.). The asymptotic confidence intervals (c.i.) are computed at the 95% level of confidence.

Dependent variable	Regressors		
		RCC	PFC
BMI	r.c.	2.286	0.013
	s.e.	0.339	0.003
	c.i.	(1.621, 2.950)	(0.007, 0.019)
SSF	r.c.	-7.746	-
	s.e.	2.783	-
	c.i.	(-13.200, -2.292)	-
PBF	r.c.	-2.724	-0.005
	s.e.	0.565	0.002
	c.i.	(-3.832, -1.616)	(-0.009, -0.001)
LBM	r.c.	14.211	0.052
	s.e.	1.649	0.015
	c.i.	(10.979, 17.442)	(0.023, 0.082)

Table 4: Joint classification of the athletes according to gender and cluster membership estimated by the best model.

Cluster	Gender		
	Female	Male	
1	39	86	125
2	61	16	77
	100	102	202

Table 5: Means and standard deviations (s.d.) of bootstrap replicates of the regression coefficients for the best model fitted to the AIS dataset. Bootstrap confidence intervals are computed at the 95% level of confidence.

Dependent variable		Regressors	
		RCC	PFC
BMI	means	2.287	0.013
	s.d.	0.327	0.003
	c.i.	(1.656, 2.932)	(0.007, 0.019)
SSF	means	-7.746	-
	s.d.	2.661	-
	c.i.	(-13.066, -2.529)	-
PBF	means	-2.724	-0.005
	s.d.	0.533	0.002
	c.i.	(-3.795, -1.697)	(-0.009, -0.001)
LBM	means	14.198	0.052
	s.d.	1.526	0.014
	c.i.	(11.222, 17.091)	(0.024, 0.080)

3.2. A numerical study of some properties of the ML estimates of β

Properties of the ML estimates of the regression coefficients are evaluated using the parametric bootstrapping residual method (Efron and Tibshirani, 1993). Namely, 5000 bootstrap samples of $I = 202$ observations each are generated from the best seemingly unrelated linear regression model described in Section 3.1 with parameters equal to the estimates provided in Tables 2 and 3. For each sample, the ML estimates of the model parameters are computed. For two bootstrap samples this computation is not performed due to the singularity of some matrices. Table 5 provides the means and standard deviations of the 4998 ML estimates of the regression coefficients as well as the bootstrap 95% confidence intervals obtained using the percentile method.

The comparison between the results in Tables 3 and 5 allows to obtain a numerical evaluation of the properties of the ML estimator for the parameters of the selected model. From the differences between the estimated regression coefficients and the means of the bootstrap replicates it emerges that the bias of the ML estimator is negligible for all regression coefficients. Namely, all the ratios between the absolute value of each bias and the bootstrap estimate for the corresponding standard error are lower than 0.04. As far as the estimates of the standard errors are concerned, the relative differences between the asymptotic and bootstrap estimates range from -3.4% (regression coefficient of PFC on PBF: 0.001863 against 0.001929) to 8.1% (regression coefficient of RCC on LBM). These differences in the estimates of the standard errors reflect upon the differences in the confidence intervals: the asymptotic confidence intervals are narrower than the bootstrap intervals when the asymptotic standard errors are smaller than the corresponding bootstrap ones. It is worth noting that the ML estimates are almost in the centre of the corresponding bootstrap confidence

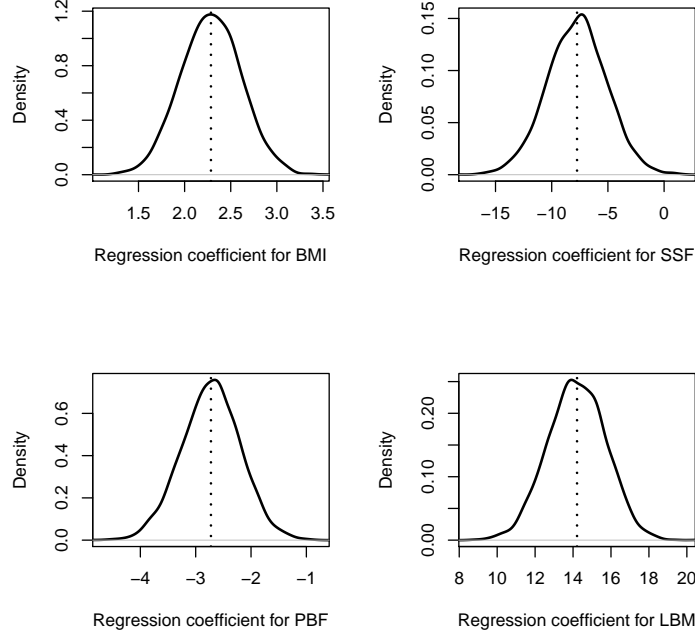


Figure 2: Estimated p.d.f. of the ML estimators of the regression coefficients of RCC on BMI, SSF, PBF and LBM, based on the bootstrap samples.

intervals. These results are related to the shape of the p.d.f. of the ML estimators. Figures 2 and 3 show the estimates of these p.d.f.'s obtained by applying the kernel method to the bootstrap replicates (the bandwidths were selected according to Sheather and Jones (1991)); ML estimates are depicted using vertical dotted lines. The distributions result to be approximately symmetric about the ML estimates.

4. Concluding remarks

In this paper, multivariate Gaussian mixtures are used to model the error terms in seemingly unrelated linear regressions. This allows to exploit the flexibility of mixtures for dealing with non Gaussian errors. In particular, the resulting models are able to handle asymmetric and heavy-tailed errors and to detect and capture the effect of relevant nominal regressors omitted from the model. Furthermore, by setting the number of components equal to one or by constraining all the equations to have the same regressors, some solutions already described in the statistical literature can be obtained as special cases.

Parsimonious seemingly unrelated linear regression models can be obtained by introducing some constraints on the component covariance matrices Σ_k 's,

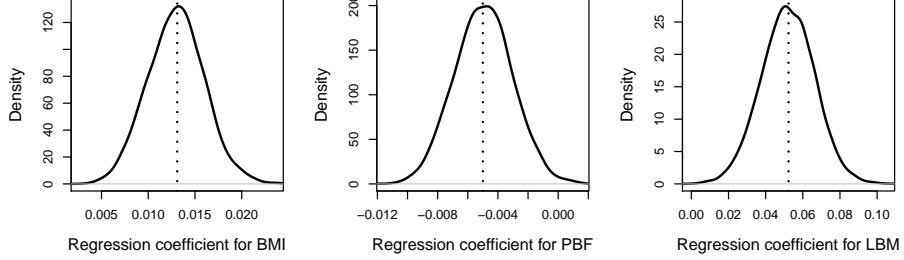


Figure 3: Estimated p.d.f. of the ML estimators of the regression coefficients of PFC on BMI, PBF and LBM, based on the bootstrap samples.

based on the spectral decomposition (see, e.g., Banfield and Raftery, 1993; Celeux and Govaert, 1995; McLachlan *et al.*, 2003; McNicholas and Murphy, 2008). Such models could provide a good fit for some datasets by using a lower number of parameters; they could be useful especially in the presence of a large number of dependent variables.

In Section 3 the *BIC* is used to select the relevant regressors in each equation as well as the number of mixture components. The use of this criterion can be motivated on the basis of both theoretical and practical results (see, e.g., Cutler and Windham, 1994; Keribin, 2000; Ray and Lindsay, 2008; Maugis *et al.*, 2009a,b). Clearly, other model selection criteria could be used, such as the *ICL* (Biernacki *et al.* (2000)), which additionally takes into account the uncertainty of the classification of the sample units to the mixture components.

Some computational issues could arise when using the models proposed in this paper. For example, when the number of candidate regressors is large, an exhaustive search for the relevant regressors for each equation could be unfeasible. A possible solution could be obtained by resorting to stochastic search techniques, such as genetic algorithms (see, e.g., Chatterjee *et al.*, 1996). As far as the EM algorithm is concerned, different initialisation strategies may be considered and evaluated (see, e.g., Biernacki *et al.*, 2003; Melnykov and Melnykov, 2012). Although these issues were not the main focus of this paper, they could deserve further investigation.

Appendix A. Proof of Theorem 2

The proof is based on the computation of the first order differential of $l(\boldsymbol{\theta})$. The model log-likelihood in equation (9) can be expressed as $l(\boldsymbol{\theta}) = \sum_{i=1}^I \ln \left(\sum_{k=1}^K f_{ki} \right)$. Thus, the first differential of $l(\boldsymbol{\theta})$ is

$$d l(\boldsymbol{\theta}) = \sum_{i=1}^I d \ln \left(\sum_{k=1}^K f_{ki} \right) = \sum_{i=1}^I \left(\sum_{k=1}^K \alpha_{ki} d \ln f_{ki} \right). \quad (\text{A.1})$$

Up to an additive constant, $\ln f_{ki}$ is equal to

$$\ln \pi_k - \frac{1}{2} \ln \det(\Sigma_k) - \frac{1}{2} \text{tr} \left[\Sigma_k^{-1} (\mathbf{y}_i - \boldsymbol{\lambda}_k - \mathbf{X}_i' \boldsymbol{\beta}) (\mathbf{y}_i - \boldsymbol{\lambda}_k - \mathbf{X}_i' \boldsymbol{\beta})' \right],$$

and

$$d \ln f_{ki} = d \ln \pi_k + d_{ki1} + d_{ki2} + d_{ki3}, \quad (\text{A.2})$$

where

$$d_{ki1} = -\frac{1}{2} d(\ln \det(\Sigma_k)), \quad (\text{A.3})$$

$$d_{ki2} = -\frac{1}{2} \text{tr} \left[d(\Sigma_k^{-1}) (\mathbf{y}_i - \boldsymbol{\lambda}_k - \mathbf{X}_i' \boldsymbol{\beta}) (\mathbf{y}_i - \boldsymbol{\lambda}_k - \mathbf{X}_i' \boldsymbol{\beta})' \right], \quad (\text{A.4})$$

$$d_{ki3} = -\frac{1}{2} \text{tr} \left[\Sigma_k^{-1} d \left((\mathbf{y}_i - \boldsymbol{\lambda}_k - \mathbf{X}_i' \boldsymbol{\beta}) (\mathbf{y}_i - \boldsymbol{\lambda}_k - \mathbf{X}_i' \boldsymbol{\beta})' \right) \right]. \quad (\text{A.5})$$

The four terms in equation (A.2) can be re-expressed as follows:

$$d \ln \pi_k = (d\boldsymbol{\pi})' \mathbf{a}_k, \quad (\text{A.6})$$

$$d_{ki1} = -\frac{1}{2} \text{tr} \left[(d\Sigma_k) \Sigma_k^{-1} \right], \quad (\text{A.7})$$

$$d_{ki2} = \frac{1}{2} \text{tr} \left[(d\Sigma_k) \mathbf{b}_{ki} \mathbf{b}_{ki}' \right], \quad (\text{A.8})$$

$$d_{ki3} = (d\boldsymbol{\lambda}_k)' \mathbf{b}_{ki} + (d\boldsymbol{\beta})' \mathbf{X}_i \mathbf{b}_{ki}, \quad (\text{A.9})$$

where equations (A.7)-(A.9) are obtained by exploiting some results from matrix derivatives (Magnus and Neudecker 1988, pgs. 182-183; Schott 2005, pgs. 292, 293, 361). Since the sum of d_{ki1} and d_{ki2} results in

$$d_{ki1} + d_{ki2} = -\frac{1}{2} d(\mathbf{v}\Sigma_k)' \mathbf{G}' \text{vec}(\mathbf{B}_{ki}), \quad (\text{A.10})$$

(see Schott, 2005, pgs. 293, 313, 356, 374), inserting equations (A.6), (A.9) and (A.10) in equation (A.2) leads to

$$\begin{aligned} d \ln f_{ki} &= (d\boldsymbol{\pi})' \mathbf{a}_k + (d\boldsymbol{\beta})' \mathbf{X}_i \mathbf{b}_{ki} + (d\boldsymbol{\lambda}_k)' \mathbf{b}_{ki} - \frac{1}{2} d(\mathbf{v}\Sigma_k)' \mathbf{G}' \text{vec}(\mathbf{B}_{ki}) \\ &= (d\boldsymbol{\pi})' \mathbf{a}_k + (d\boldsymbol{\beta})' \mathbf{X}_i \mathbf{b}_{ki} + (d\boldsymbol{\theta}_k)' \mathbf{c}_{ki}. \end{aligned} \quad (\text{A.11})$$

Using equations (A.1) and (A.11), $dl(\boldsymbol{\theta})$ can be expressed as

$$dl(\boldsymbol{\theta}) = (d\boldsymbol{\pi})' \sum_{i=1}^I \sum_{k=1}^K \alpha_{ki} \mathbf{a}_k + (d\boldsymbol{\beta})' \sum_{i=1}^I \mathbf{X}_i \sum_{k=1}^K \alpha_{ki} \mathbf{b}_{ki} + \sum_{k=1}^K (d\boldsymbol{\theta}_k)' \sum_{i=1}^I \alpha_{ki} \mathbf{c}_{ki}, \quad (\text{A.12})$$

thus proving the theorem.

Appendix B. Proof of Theorem 3

The proof is based on the computation of the second order differential of $l(\boldsymbol{\theta})$:

$$d^2 l(\boldsymbol{\theta}) = \sum_{i=1}^I d^2 \ln \left(\sum_{k=1}^K f_{ki} \right), \quad (\text{B.1})$$

where

$$d^2 \ln \left(\sum_{k=1}^K f_{ki} \right) = \sum_{k=1}^K \alpha_{ki} d^2 \ln f_{ki} + \sum_{k=1}^K \alpha_{ki} (d \ln f_{ki})^2 - \left(\sum_{k=1}^K \alpha_{ki} d \ln f_{ki} \right)^2 \quad (\text{B.2})$$

(see Boldea and Magnus, 2009, Appendix).

Since $(d \ln f_{ki})^2 = (d \ln f_{ki}) (d \ln f_{ki})'$, using equation (A.11) it results that

$$\begin{aligned} (d \ln f_{ki})^2 &= (d\boldsymbol{\pi})' \mathbf{a}_k \mathbf{a}_k' d\boldsymbol{\pi} + (d\boldsymbol{\pi})' \mathbf{a}_k \mathbf{b}_{ki}' \mathbf{X}_i' d\boldsymbol{\beta} + (d\boldsymbol{\pi})' \mathbf{a}_k \mathbf{c}_{ki}' d\boldsymbol{\theta}_k \\ &\quad + (d\boldsymbol{\beta})' \mathbf{X}_i \mathbf{b}_{ki} \mathbf{a}_k' d\boldsymbol{\pi} + (d\boldsymbol{\beta})' \mathbf{X}_i \mathbf{b}_{ki} \mathbf{b}_{ki}' \mathbf{X}_i' d\boldsymbol{\beta} + (d\boldsymbol{\beta})' \mathbf{X}_i \mathbf{b}_{ki} \mathbf{c}_{ki}' d\boldsymbol{\theta}_k \\ &\quad + (d\boldsymbol{\theta}_k)' \mathbf{c}_{ki} \mathbf{a}_k' d\boldsymbol{\pi} + (d\boldsymbol{\theta}_k)' \mathbf{c}_{ki} \mathbf{b}_{ki}' \mathbf{X}_i' d\boldsymbol{\beta} + (d\boldsymbol{\theta}_k)' \mathbf{c}_{ki} \mathbf{c}_{ki}' d\boldsymbol{\theta}_k. \end{aligned} \quad (\text{B.3})$$

Similarly,

$$\begin{aligned} \left(\sum_{k=1}^K \alpha_{ki} d \ln f_{ki} \right)^2 &= \left(\sum_{k=1}^K \alpha_{ki} d \ln f_{ki} \right) \left(\sum_{k=1}^K \alpha_{ki} d \ln f_{ki} \right)' \\ &= (d\boldsymbol{\pi})' \bar{\mathbf{a}}_i \bar{\mathbf{a}}_i' d\boldsymbol{\pi} + (d\boldsymbol{\pi})' \bar{\mathbf{a}}_i \bar{\mathbf{b}}_i' \mathbf{X}_i' d\boldsymbol{\beta} + (d\boldsymbol{\pi})' \bar{\mathbf{a}}_i \sum_{k=1}^K \alpha_{ki} \mathbf{c}_{ki}' d\boldsymbol{\theta}_k \\ &\quad + (d\boldsymbol{\beta})' \mathbf{X}_i \bar{\mathbf{b}}_i \bar{\mathbf{a}}_i' d\boldsymbol{\pi} + (d\boldsymbol{\beta})' \mathbf{X}_i \bar{\mathbf{b}}_i \bar{\mathbf{b}}_i' \mathbf{X}_i' d\boldsymbol{\beta} \\ &\quad + (d\boldsymbol{\beta})' \mathbf{X}_i \bar{\mathbf{b}}_i \sum_{k=1}^K \alpha_{ki} \mathbf{c}_{ki}' d\boldsymbol{\theta}_k + \left[\sum_{k=1}^K (d\boldsymbol{\theta}_k)' \alpha_{ki} \mathbf{c}_{ki} \right] \bar{\mathbf{a}}_i' d\boldsymbol{\pi} \\ &\quad + \left[\sum_{k=1}^K (d\boldsymbol{\theta}_k)' \alpha_{ki} \mathbf{c}_{ki} \right] \bar{\mathbf{b}}_i' \mathbf{X}_i' d\boldsymbol{\beta} \\ &\quad + \sum_{k=1}^K \sum_{h=1}^K (d\boldsymbol{\theta}_k)' \alpha_{ki} \alpha_{hi} \mathbf{c}_{ki} \mathbf{c}_{hi}' d\boldsymbol{\theta}_l. \end{aligned} \quad (\text{B.4})$$

Furthermore,

$$\begin{aligned} d^2 \ln f_{ki} &= - (d\boldsymbol{\pi})' \mathbf{a}_k \mathbf{a}_k' d\boldsymbol{\pi} - (d\boldsymbol{\beta})' \mathbf{X}_i \boldsymbol{\Sigma}_k^{-1} \mathbf{X}_i' d\boldsymbol{\beta} - (d\boldsymbol{\theta}_k)' \mathbf{F}_{ki}' \mathbf{X}_i' d\boldsymbol{\beta} \\ &\quad - (d\boldsymbol{\beta})' \mathbf{X}_i \mathbf{F}_{ki} d\boldsymbol{\theta}_k - (d\boldsymbol{\theta}_k)' \mathbf{C}_{ki} d\boldsymbol{\theta}_k \end{aligned} \quad (\text{B.5})$$

(see Appendix C). From equations (B.2), (B.3), (B.4) and (C.1) and by group-

ing together the common factors it follows that

$$\begin{aligned}
d^2 \ln \left(\sum_{k=1}^K f_{ki} \right) = & - (d\boldsymbol{\pi})' \bar{\mathbf{a}}_i \bar{\mathbf{a}}_i' d\boldsymbol{\pi} + (d\boldsymbol{\pi})' \left[\left(\sum_{k=1}^K \alpha_{ki} \mathbf{a}_k \mathbf{b}_{ki}' \right) - \bar{\mathbf{a}}_i \bar{\mathbf{b}}_i' \right] \mathbf{X}_i' d\boldsymbol{\beta} \\
& + (d\boldsymbol{\pi})' \left[\sum_{k=1}^K \alpha_{ki} (\mathbf{a}_k - \bar{\mathbf{a}}_i) \mathbf{c}_{ki}' d\boldsymbol{\theta}_k \right] \\
& + (d\boldsymbol{\beta})' \mathbf{X}_i \left[\left(\sum_{k=1}^K \alpha_{ki} \mathbf{b}_{ki} \mathbf{a}_k' \right) - \bar{\mathbf{b}}_i \bar{\mathbf{a}}_i' \right] d\boldsymbol{\pi} \\
& - (d\boldsymbol{\beta})' \mathbf{X}_i [\bar{\mathbf{B}}_i + \bar{\mathbf{b}}_i \bar{\mathbf{b}}_i'] \mathbf{X}_i' d\boldsymbol{\beta} \\
& - (d\boldsymbol{\beta})' \mathbf{X}_i \left\{ \sum_{k=1}^K \alpha_{ki} [\mathbf{F}_{ki} - (\mathbf{b}_{ki} - \bar{\mathbf{b}}_i) \mathbf{c}_{ki}'] d\boldsymbol{\theta}_k \right\} \\
& + \left[\sum_{k=1}^K (d\boldsymbol{\theta}_k)' \alpha_{ki} \mathbf{c}_{ki} (\mathbf{a}_k' - \bar{\mathbf{a}}_i') \right] d\boldsymbol{\pi} \\
& - \left\{ \sum_{k=1}^K (d\boldsymbol{\theta}_k)' \alpha_{ki} [\mathbf{F}_{ki}' - \mathbf{c}_{ki} (\mathbf{b}_{ki}' - \bar{\mathbf{b}}_i')] \right\} \mathbf{X}_i' d\boldsymbol{\beta} \\
& - \sum_{k=1}^K (d\boldsymbol{\theta}_k)' \alpha_{ki} [\mathbf{C}_{ki} - \mathbf{c}_{ki} \mathbf{c}_{ki}'] d\boldsymbol{\theta}_k \\
& - \sum_{k=1}^K \sum_{h=1}^K [(d\boldsymbol{\theta}_k)' \alpha_{ki} \alpha_{hi} \mathbf{c}_{ki} \mathbf{c}_{hi}' d\boldsymbol{\theta}_h] . \tag{B.6}
\end{aligned}$$

Inserting equation (B.6) in equation (B.1) completes the proof.

Appendix C. Second order differential of $\ln f_{ki}$

Using equation (A.2) the second order differential of $\ln f_{ki}$ can be expressed as

$$d^2 \ln f_{ki} = d^2 \ln \pi_k + d(d_{ki1}) + d(d_{ki2}) + d(d_{ki3}). \tag{C.1}$$

From equation (A.6) it follows that

$$d^2 \ln \pi_k = - (d\boldsymbol{\pi})' \mathbf{a}_k \mathbf{a}_k' d\boldsymbol{\pi}. \tag{C.2}$$

The second term in equation (C.1) is equal to

$$d(d_{ki1}) = -\frac{1}{2} \text{tr} [d\boldsymbol{\Sigma}_k (d\boldsymbol{\Sigma}_k^{-1})] = \frac{1}{2} \text{tr} [(d\boldsymbol{\Sigma}_k) \boldsymbol{\Sigma}_k^{-1} (d\boldsymbol{\Sigma}_k) \boldsymbol{\Sigma}_k^{-1}]. \tag{C.3}$$

The third term that composes $d^2 \ln f_{ki}$ results to be

$$\begin{aligned} d(d_{ki2}) &= \frac{1}{2} \text{tr} \left[d(\Sigma_k^{-1}) (d\Sigma_k) \Sigma_k^{-1} (\mathbf{y}_i - \boldsymbol{\lambda}_k - \mathbf{X}'_i \boldsymbol{\beta}) (\mathbf{y}_i - \boldsymbol{\lambda}_k - \mathbf{X}'_i \boldsymbol{\beta})' \right] \\ &\quad + \frac{1}{2} \text{tr} \left[\Sigma_k^{-1} (d\Sigma_k) d(\Sigma_k^{-1}) (\mathbf{y}_i - \boldsymbol{\lambda}_k - \mathbf{X}'_i \boldsymbol{\beta}) (\mathbf{y}_i - \boldsymbol{\lambda}_k - \mathbf{X}'_i \boldsymbol{\beta})' \right] \\ &\quad + \frac{1}{2} \text{tr} \left[\Sigma_k^{-1} (d\Sigma_k) \Sigma_k^{-1} d \left((\mathbf{y}_i - \boldsymbol{\lambda}_k - \mathbf{X}'_i \boldsymbol{\beta}) (\mathbf{y}_i - \boldsymbol{\lambda}_k - \mathbf{X}'_i \boldsymbol{\beta})' \right) \right]. \end{aligned}$$

By exploiting some properties of the trace of a square matrix (see, e.g. Schott, 2005), $d(d_{ki2})$ can also be expressed as

$$\begin{aligned} d(d_{ki2}) &= \text{tr} \left[(d\Sigma_k) d(\Sigma_k^{-1}) (\mathbf{y}_i - \boldsymbol{\lambda}_k - \mathbf{X}'_i \boldsymbol{\beta}) (\mathbf{y}_i - \boldsymbol{\lambda}_k - \mathbf{X}'_i \boldsymbol{\beta})' \Sigma_k^{-1} \right] \\ &\quad + \frac{1}{2} \text{tr} \left[\Sigma_k^{-1} (d\Sigma_k) \Sigma_k^{-1} d \left((\mathbf{y}_i - \boldsymbol{\lambda}_k - \mathbf{X}'_i \boldsymbol{\beta}) (\mathbf{y}_i - \boldsymbol{\lambda}_k - \mathbf{X}'_i \boldsymbol{\beta})' \right) \right], \end{aligned}$$

and using two theorems about the vec and trace operators (Schott, 2005, Theorems 8.9 and 8.12) it follows that

$$\begin{aligned} d(d_{ki2}) &= \text{tr} \left[(d\Sigma_k) d(\Sigma_k^{-1}) (\mathbf{y}_i - \boldsymbol{\lambda}_k - \mathbf{X}'_i \boldsymbol{\beta}) (\mathbf{y}_i - \boldsymbol{\lambda}_k - \mathbf{X}'_i \boldsymbol{\beta})' \Sigma_k^{-1} \right] \\ &\quad - (d\boldsymbol{\lambda}_k)' (\mathbf{b}'_{ki} \otimes \Sigma_k^{-1}) d(\text{vec} \Sigma_k) \\ &\quad - (d\boldsymbol{\beta})' \mathbf{X}_i (\mathbf{b}'_{ki} \otimes \Sigma_k^{-1}) d(\text{vec} \Sigma_k). \end{aligned} \quad (\text{C.4})$$

From equations (C.3) and (C.4) it follows that

$$\begin{aligned} d(d_{ki1}) + d(d_{ki2}) &= \frac{1}{2} \text{tr} \left[(d\Sigma_k) \Sigma_k^{-1} (d\Sigma_k) \Sigma_k^{-1} \right] \\ &\quad - \text{tr} \left[(d\Sigma_k) \Sigma_k^{-1} (d\Sigma_k) \mathbf{b}_{ki} \mathbf{b}'_{ki} \right] \\ &\quad - (d\boldsymbol{\lambda}_k)' (\mathbf{b}'_{ki} \otimes \Sigma_k^{-1}) d(\text{vec} \Sigma_k) \\ &\quad - (d\boldsymbol{\beta})' \mathbf{X}_i (\mathbf{b}'_{ki} \otimes \Sigma_k^{-1}) d(\text{vec} \Sigma_k) \\ &= \frac{1}{2} \text{tr} \left\{ (d\Sigma_k) \Sigma_k^{-1} (d\Sigma_k) [\Sigma_k^{-1} + \Sigma_k^{-1} - \Sigma_k^{-1} - 2\mathbf{b}_{ki} \mathbf{b}'_{ki}] \right\} \\ &\quad - (d\boldsymbol{\lambda}_k)' (\mathbf{b}'_{ki} \otimes \Sigma_k^{-1}) d(\text{vec} \Sigma_k) \\ &\quad - (d\boldsymbol{\beta})' \mathbf{X}_i (\mathbf{b}'_{ki} \otimes \Sigma_k^{-1}) d(\text{vec} \Sigma_k) \\ &= -\frac{1}{2} \text{vec} \left((d\Sigma_k)' \right)' \left[(\Sigma_k^{-1} - 2\mathbf{B}_{ki})' \otimes \Sigma_k^{-1} \right] \text{vec} (d\Sigma_k) \\ &\quad - (d\boldsymbol{\lambda}_k)' (\mathbf{b}'_{ki} \otimes \Sigma_k^{-1}) d(\text{vec} \Sigma_k) \\ &\quad - (d\boldsymbol{\beta})' \mathbf{X}_i (\mathbf{b}'_{ki} \otimes \Sigma_k^{-1}) d(\text{vec} \Sigma_k) \\ &= -\frac{1}{2} d(\mathbf{v} \Sigma_k)' \mathbf{G}' \left[(\Sigma_k^{-1} - 2\mathbf{B}_{ki}) \otimes \Sigma_k^{-1} \right] \mathbf{G} d(\mathbf{v} \Sigma_k) \\ &\quad - (d\boldsymbol{\lambda}_k)' (\mathbf{b}'_{ki} \otimes \Sigma_k^{-1}) \mathbf{G} d(\mathbf{v} \Sigma_k) \\ &\quad - (d\boldsymbol{\beta})' \mathbf{X}_i (\mathbf{b}'_{ki} \otimes \Sigma_k^{-1}) \mathbf{G} d(\mathbf{v} \Sigma_k), \end{aligned} \quad (\text{C.5})$$

where the third and fourth equalities are obtained using some properties of the vec operator (see Schott, 2005, pg. 294).

From equation (A.9) it is possible to write

$$\begin{aligned}
d(d_{ki3}) &= (d\lambda_k)' d\mathbf{b}_{ki} + (d\beta)' \mathbf{X}_i d\mathbf{b}_{ki} \\
&= - (d\lambda_k)' \Sigma_k^{-1} d(\Sigma_k) \mathbf{b}_{ki} - (d\lambda_k)' \Sigma_k^{-1} d\lambda_k - (d\lambda_k)' \Sigma_k^{-1} \mathbf{X}_i' d\beta \\
&\quad - (d\beta)' \mathbf{X}_i \Sigma_k^{-1} d(\Sigma_k) \mathbf{b}_{ki} - (d\beta)' \mathbf{X}_i \Sigma_k^{-1} d\lambda_k - (d\beta)' \mathbf{X}_i \Sigma_k^{-1} \mathbf{X}_i' d\beta \\
&= -d(\mathbf{v}\Sigma_k)' \mathbf{G}' (\mathbf{b}_{ki} \otimes \Sigma_k^{-1}) d\lambda_k - (d\lambda_k)' \Sigma_k^{-1} d\lambda_k \\
&\quad - (d\lambda_k)' \Sigma_k^{-1} \mathbf{X}_i' d\beta - d(\mathbf{v}\Sigma_k)' \mathbf{G}' (\mathbf{b}_{ki} \otimes \Sigma_k^{-1}) \mathbf{X}_i' d\beta \\
&\quad - (d\beta)' \mathbf{X}_i \Sigma_k^{-1} d\lambda_k - (d\beta)' \mathbf{X}_i \Sigma_k^{-1} \mathbf{X}_i' d\beta, \tag{C.6}
\end{aligned}$$

where the third equality results from the same theorems about the vec and trace operators employed above and the second equality is obtained using the following expression for $d\mathbf{b}_{ki}$:

$$\begin{aligned}
d\mathbf{b}_{ki} &= d(\Sigma_k^{-1}) (\mathbf{y}_i - \lambda_k - \mathbf{X}_i' \beta) + \Sigma_k^{-1} d(\mathbf{y}_i - \lambda_k - \mathbf{X}_i' \beta) \\
&= -\Sigma_k^{-1} d(\Sigma_k) \mathbf{b}_{ki} - \Sigma_k^{-1} d\lambda_k - \Sigma_k^{-1} \mathbf{X}_i' d\beta.
\end{aligned}$$

Inserting equations (C.2), (C.5) and (C.6) in equation (C.1) and using the definitions of θ_k , \mathbf{F}_{ki} and \mathbf{C}_{ki} introduced in Section 2.3 results in the following expression for $d^2 \ln f_{ki}$:

$$\begin{aligned}
d^2 \ln f_{ki} &= - (d\pi)' \mathbf{a}_k \mathbf{a}_k' d\pi - (d\beta)' \mathbf{X}_i \Sigma_k^{-1} \mathbf{X}_i' d\beta - (d\theta_k)' \mathbf{F}_{ki}' \mathbf{X}_i' d\beta \\
&\quad - (d\beta)' \mathbf{X}_i \mathbf{F}_{ki} d\theta_k - (d\theta_k)' \mathbf{C}_{ki} d\theta_k.
\end{aligned}$$

References

- Ando, T., Zellner, A.: Hierarchical Bayesian analysis of the seemingly unrelated regression and simultaneous equations models using a combination of direct Monte Carlo and importance sampling techniques. *Bayesian Anal.* **5**, 65–96 (2010)
- Banfield, J.D., Raftery, A.E.: Model-based Gaussian and non-Gaussian clustering. *Biometrics* **49**, 803–821 (1993)
- Bartolucci, F., Scaccia, L.: The use of mixtures for dealing with non-normal regression errors. *Comput. Stat. Data Anal.* **48**, 821–834 (2005)
- Biernacki, C., Celeux, G., Govaert, G.: Assessing a mixture model for clustering with the integrated classification likelihood. *IEEE Trans. Pattern Anal. Mach. Intell.* **22**, 719–725 (2000)
- Biernacki, C., Celeux, G., Govaert, G.: Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate Gaussian mixture models. *Comput. Stat. Data Anal.* **41**, 561–575 (2003)
- Boldea, O., Magnus, J.R.: Maximum likelihood estimation of the multivariate normal mixture model. *J. Am. Stat. Assoc.* **104**, 1539–1549 (2009)

- Chatterjee, S., Laudato, M., Lynch, L.A.: Genetic algorithms and their statistical applications: an introduction. *Comput. Stat. Data Anal.* **22**, 633–651 (1996)
- Celeux, G., Govaert, G.: Gaussian parsimonious clustering models. *Pattern Recognit.* **28**, 781–793 (1995)
- Cook, R.D., Weisberg, S.: *An Introduction to Regression Graphics*. Wiley, New York (1994)
- Cutler, A., Windham, M.P.: Information-based validity functionals for mixture analysis. In: Bozdogan, H. (ed.) *Proceedings of the First US/Japan Conference on the Frontiers of Statistical Modeling: An Informational Approach*, pp. 149–170. Kluwer Academic, Dordrecht (1994)
- Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood for incomplete data via the EM algorithm. *J. R. Stat. Soc. Ser. B* **39**, 1–22 (1977)
- Efron, B., Tibshirani, R.J.: *An Introduction to the Bootstrap*. Chapman & Hall, London (1993)
- Fraser, D.A.S., Rekkas, M., Wong, A.: Highly accurate likelihood analysis for the seemingly unrelated regression problem. *J. Econom.* **127**, 17–33 (2005)
- Galimberti, G., Soffritti, G.: A multivariate linear regression analysis using finite mixtures of t distributions. *Comput. Stat. Data Anal.* **71**, 138–150 (2014)
- Keribin, C.: Consistent estimation of the order of mixture models. *Sankhyā Ser. A*, **62**, 49–66 (2000)
- Kmenta, J., Gilbert, R.: Small sample properties of alternative estimators of seemingly unrelated regressions. *J. Am. Stat. Assoc.* **63**, 1180–1200 (1968)
- Kowalski, J., Mendoza-Blanco, J.R., Tu, X.M., Gleser, L.J.: On the difference in inference and prediction between the joint and independent t -error models for seemingly unrelated regressions. *Commun. Stat. Theory* **28**, 2119–2140 (1999)
- Kurata, H.: On the efficiencies of several generalized least squares estimators in a seemingly unrelated regression model and a heteroscedastic model. *J. Multivar. Anal.* **70**, 86–94 (1999)
- Magnus, J.R.: Maximum likelihood estimation of the GLS model with unknown parameters in the disturbance covariance matrix. *J. Econom.* **7**, 281–312 (1978)
- Magnus, J.R., Neudecker, H.: *Matrix Differential Calculus with Applications in Statistics and Econometrics*. John Wiley & Sons, Chichester (1988)

- Maugis, C., Celeux, G., Martin-Magniette, M.-L.: Variable selection in model-based clustering: a general variable role modeling. *Comput. Stat. Data Anal.* **53**, 3872–3882 (2009a)
- Maugis, C., Celeux, G., Martin-Magniette, M.-L.: Variable selection for clustering with Gaussian mixture models. *Biometrics* **65**, 701–709 (2009b)
- McLachlan, G.J., Krishnan, T.: *The EM Algorithm and Extensions*. 2nd edn. Wiley, Chichester (2008)
- McLachlan, G.J., Peel, D.: *Finite Mixture Models*. Wiley, Chichester (2000)
- McLachlan, G.J., Peel, D., Bean, R.W.: Modelling high-dimensional data by mixtures of factor analyzers. *Comput. Stat. Data Anal.* **41**, 379–388 (2003)
- McNicholas, P.D., Murphy, T.B.: Parsimonious Gaussian mixture models. *Stat. Comput.* **18**, 285–296 (2008)
- Melnykov, V., Melnykov, I.: Initializing the EM algorithm in Gaussian mixture models with an unknown number of components. *Comput. Stat. Data Anal.* **56**, 1381–1395 (2012)
- Ng, V.M.: Robust Bayesian inference for seemingly unrelated regressions with elliptical errors. *J. Multivar. Anal.* **83**, 409–414 (2002)
- Oberhofer, W., Kmenta, J.: A general procedure for obtaining maximum likelihood estimates in generalized regression models. *Econometrica* **42**, 579–590 (1974)
- Park, T.: Equivalence of maximum likelihood estimation and iterative two-stage estimation for seemingly unrelated regression models. *Commun. Stat. Theory* **22**, 2285–2296 (1993)
- Percy, D.F.: Predictions for seemingly unrelated regression. *J. R. Stat. Soc. Ser. B* **54**, 243–252 (1992)
- Ray, S., Lindsay, B.G.: Model selection in high dimensions: a quadratic-risk-based approach. *J. R. Stat. Soc. Ser. B* **70**, 95–118 (2008)
- R Core Team: *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL <http://www.R-project.org/>(2013)
- Redner, R.A., Walker, H.F.: Mixture densities, maximum likelihood and the EM algorithm. *SIAM Rev.* **26**, 195–239 (1984)
- Rilstone, P., Veall, M.: Using bootstrapped confidence intervals for improved inferences with seemingly unrelated regression equations. *Econom. Theory* **12**, 569–580 (1996)

- Rocke, D.: Bootstrap Bartlett adjustment in seemingly unrelated regression. *J. Am. Stat. Assoc.* **84**, 598–601 (1989)
- Schott, J.R.: *Matrix Analysis for Statistics*. 2nd edn. John Wiley & Sons, New York (2005)
- Schwarz, G.: Estimating the dimension of a model. *Ann. Stat.* **6**, 461–464 (1978)
- Sheather, S.J., Jones, M.C.: A reliable data-based bandwidth selection method for kernel density estimation. *J. R. Stat. Soc. Ser. B* **53**, 683–690 (1991)
- Soffritti, G., Galimberti, G.: Multivariate linear regression with non-normal errors: a solution based on mixture models. *Stat. Comput.* **21**, 523–536 (2011)
- Srivastava, V.K., Giles, D.E.A.: *Seemingly Unrelated Regression Equations Models*. Marcel Dekker, New York (1987)
- Srivastava, V.K., Maekawa, K.: Efficiency properties of feasible generalized least squares estimators in SURE models under non-normal disturbances. *J. Econom.* **66**, 99–121 (1995)
- Zellner, A.: An efficient method of estimating seemingly unrelated regression equations and tests for aggregation bias. *J. Am. Stat. Assoc.* **57**, 348–368 (1962)
- Zellner, A.: Estimators for seemingly unrelated regression equations: some exact finite sample results. *J. Am. Stat. Assoc.* **58**, 977–992 (1963)
- Zellner, A.: *An Introduction to Bayesian Inference in Econometrics*. Wiley, New York (1971)
- Zellner, A., Ando, T.: A direct Monte Carlo approach for Bayesian analysis of the seemingly unrelated regression model. *J. Econom.* **159**, 33–45 (2010a)
- Zellner, A., Ando, T.: Bayesian and non-Bayesian analysis of the seemingly unrelated regression model with Student- t errors, and its application for forecasting. *Int. J. Forecast.* **26**, 413–434 (2010b)